

Илья Дуванов

Информация о выбранном датасете

Обновлено 22 мая 2025, 16:36

Содержание страницы

Общая информация

- [Описание данных](#)

- [Структуры данных](#)

- [Поля в hour.csv](#)

- [Поля в day.csv](#)

- [Статистические характеристики](#)

- [Общая статистика \(hour.csv\)](#)

- [Распределение по времени года](#)

- [Пропущенные значения](#)

- [Использование в проекте](#)

- [Цель использования](#)

- [Подготовка данных](#)

- [Ограничения датасета](#)

Общая информация

UCI Bike Sharing Dataset — это общедоступный набор данных, предоставленный сообществом UCI Machine Learning Repository. Датасет содержит данные о прокате велосипедов в Вашингтоне, округ Колумбия, собранные с 1 января 2011 года по 31 декабря 2012 года. Он широко используется для задач машинного обучения, связанных с прогнозированием спроса, временных рядов и анализа влияния погодных условий на поведение пользователей.

- **Источник:** [UCI Machine Learning Repository](#)
- **Авторы:** Hadi Fanaee-T и João Gama, LIAAD, INESC Porto, Университет Порту, Португалия.
- **Дата публикации:** 20 января 2013 года.
- **Лицензия:** Открытый доступ (без ограничений на использование для образовательных и исследовательских целей).

Описание данных

Датасет включает два основных файла:

1. **Hourly data (hour.csv):** Содержит записи с почасовым разрешением (17 379 строк).
2. **Daily data (day.csv):** Содержит записи с суточным разрешением (731 строки).

Проект использует **hour.csv**, так как задача требует прогноза спроса на ближайший час.

Структуры данных

Поля в hour.csv

Название поля	Описание	Тип данных	Диапазон значений
instant	Индекс записи (уникальный идентификатор).	Integer	1–17 379
dteday	Дата (гггг-мм-дд).	Integer	2011-01-01 – 2012-12-31
season	Время года (1: весна, 2: лето, 3: осень, 4: зима).	Date	1–4
yr	Год (0: 2011, 1: 2012).	Integer	0–1
mnth	Месяц (1–12).	Integer	1–12
hr	Час дня (0–23).	Integer	0–23
holiday	Признак праздника (0: нет, 1: да).	Integer	0–1
weekday	День недели (0: воскресенье, 1–6: понедельник–суббота).	Integer	0–6
workingday	Рабочий день (0: нет, 1: да).	Integer	0–1
weathersit	Погода (1: чисто, 2: облачно, 3: дождь, 4: сильный дождь).	Integer	1–4

temp	Нормализованная температура (в °C, от 0 до 1)	Float	0–1 (реальные: -8–39 °C)
atemp	Нормализованная "ощущаемая" температура.	Float	0–1
hum	Нормализованная влажность (от 0 до 1).	Float	0–1
windspeed	Нормализованная скорость ветра (от 0 до 1).	Float	0–1
casual	Количество аренд случайными пользователями.	Integer	0–694
registered	Количество аренд зарегистрированными пользователями.	Integer	0–886
cnt	Общее количество аренд (casual + registered).	Integer	1–977

Поля в day.csv

(Используются косвенно для анализа трендов, но не для почасового прогноза):

- instant, dteday, season, yr, mnth, holiday, weekday, workingday, weathersit, temp, atemp, hum, windspeed, casual, registered, cnt.

Статистические характеристики

Общая статистика (hour.csv)

Параметр	Среднее значение	Минимальное значение	Максимальное значение	Стандартное отклонение
temp	0.497	0.02	1.0	0.183
atemp	0.476	0.0	1.0	0.174
hum	0.627	0.0	1.0	0.192
windspeed	0.189	0.0	0.851	0.122
casual	35.676	0	694	49.305
registered	153.786	0	886	151.357
cnt	189.463	1	977	181.387

Распределение по времени года

- Весна (season=1): ~25% данных.
- Лето (season=2): ~25% данных.
- Осень (season=3): ~25% данных.
- Зима (season=4): ~25% данных.

Пропущенные значения

- Датасет не содержит пропущенных значений, что делает его готовым к использованию без предварительной очистки.

Использование в проекте

Цель использования

Датасет используется для разработки системы прогнозирования спроса на аренду велосипедов на основе почасовых данных. Основная задача — предсказать значение `cnt` (общее количество аренд) для следующего часа с учетом входных признаков (`hr`, `temp`, `weathersit` и др.).

Подготовка данных

- Фильтрация:** Исключение полей `casual` и `registered` из входных признаков, так как они не доступны на момент прогноза (используются только для валидации).
- Нормализация:** Все числовые признаки (`temp`, `atemp`, `hum`, `windspeed`) уже нормализованы в диапазоне [0, 1].
- Кодирование:** Категориальные признаки (`season`, `weathersit`, `weekday`) преобразуются в one-hot encoding для использования в моделях.
- Разделение:** Данные делятся на обучающую (80%) и тестовую (20%) выборки по времени.

Ограничения датасета

- Исторический характер:** Данные охватывают только 2011–2012 годы, что может не отражать текущие тенденции (например, рост популярности велопроката в 2025 году).
- Географическая привязка:** Данные специфичны для Вашингтона, округ Колумбия, и могут не подходить для других регионов без корректировки.
- Отсутствие внешних факторов:** Не учитываются такие факторы, как акции, ремонт станций или аварии, которые могут влиять на спрос.

